DOCUMENT RESUME

ED 151 283 SO 010 741

AUTHOR Bash, Maurice J.; Rasher, Sue Pinzur
TITLE Improving Local Curriculum Adoption Decision Making through Use of Criterion-Referenced Evaluation: A Case Study in Social Studies.
PUB DATE Mar 78
NOTE 29p.; Paper presented at Annual Meeting of the American Educational Research Association (Toronto, Ontario, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; *Case Studies (Education); Course Objectives; Criterion Referenced Tests; Curriculum Development; *Curriculum Evaluation; Curriculum Planning; Curriculum Research; *Decision Making; Educational Research; Elementary Education; Elementary School Curriculum; *Evaluation Methods; Field Studies; Group Tests; Inservice Courses; Instructional Materials; Item Analysis; Mastery Tests; Norm Referenced Tests; Selection; *Social Studies; *Textbook Selection

ABSTRACT

This case study recounts the attempts of two school districts to improve decision making on the adoption of a new elementary social studies curriculum by using formal evaluation methodology. The study's main objective was to develop a series of criterion referenced (mastery level) tests from a list of social studies objectives to determine whether these objectives were taught more effectively by the new curriculum. Other objectives included: developing a firm data base for decision making; involving teachers and principals in a field test to assess curriculum objectives; and projecting a plan for an inservice program. Drawing on the technical resources of the University of Illinois, the districts conducted a field evaluation of the new curriculum. Two groups of students were administered the tests; the experimental group received instruction in the new curriculum, while the control group was exposed to the more traditional curriculum. A total of 1086 students in 48 classrooms participated. Findings indicated the experimental group consistently made higher gains in achievement than the control group. It was also concluded that although criterion referenced evaluation provided important information, traditional test analysis measures proved to be more useful for making decisions on adoption and for planning teacher inservice programs. (Author/JK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151283

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Maurice Eash and Sue Rasher

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Improving Local Curriculum Adoption Decision Making Through Use of Criterion-Referenced Evaluation: A Case Study in Social Studies

Maurice J. Eash and Sue Pinzur Rasher

University of Illinois at Chicago Circle

Office of Evaluation Research

College of Education

Box 4348

Chicago, IL 60680

SP 010 741

Paper presented at the AERA Annual Meeting

Toronto, Ontario, Canada

March 27-31, 1978

Session Number 8.22

Introduction

This case study recounts the attempt of two school districts to improve decision making on the adoption of a new curriculum (grades 1-6) using formal evaluation methodology. Drawing on the technical resources of a university, they conducted a field evaluation of a new social studies curriculum based on a set of criterion-referenced tests which were cooperatively developed by a committee from the school districts and the university. The committee from the school districts was interested in examining whether the curriculum teaches the complex objectives listed by the curriculum developers (which were based in the discipline of economics). As one aspect of the evaluation, the committee set as a priority the development of a set of criterion-referenced tests which would be used in a pre-post test design to measure achievement gains.[1] Two groups of students were administered the tests; the experimental group received instruction in the curriculum,while the control group was exposed to a more traditional curriculum. This case study is not an attempt to resolve the issues in the use of criterion-referenced test (CRT) versus norn-referenced testing (NRT); however, it illustrates how a number of these issues insistently came to the fore as the decision-making process on the curriculum advanced toward closure. More specifically, early in the study, the committee adopted a stance based on an assumption that criterion-referenced testing had marked

---

1. Norm referenced tests have traditionally been recommended for comparing the effectiveness of two or more curricula, however, the districts wanted to compare curricula and measure mastery of concepts of the new curriculum. Lack of time and resources precluded the development and administration of both criterion-referenced and norm referenced tests. Thus, the districts had to choose which kind of test should be developed and administered.

advantages over norm-referenced measures. In particular, the committee felt the use of criterion-referenced testing would avoid teacher objections to comparisons of student achievement among classrooms. Thus, the committee would meet the growing demand for accountability in selection and use of materials by embarking on a revised approach to adopting a new curriculum: an empirical testing of the curriculum before investing in a system-wide adoption.

The objectives of the evaluative study were:

1. To develop a firmer data base for decision-making in adopting new curriculum in the two school districts.
2. To develop a series of criterion-referenced tests over a list of social studies objectives to determine whether these objectives were taught more effectively by the new curriculum.
3. To involve a group of teachers and principals in a field test of the new curriculum as an inservice effort to assess the objectives and direction of the social studies curriculum.
4. To project a plan for a continuing inservice program on the social studies curriculum from the findings of the evaluative study.

Instructional materials represent a major time commitment for students since up to 80 percent of a student's classroom hours are spent engaging materials (EPIE, 1977a). Despite this known time commitment and the increasing awareness of the importance of the time variable in learning (Carroll, 1963), the selection and adoption of instructional materials has been haphazard in many school districts, with teachers spending as little as one hour per school year in the process. Nevertheless as the major foci of the curriculum which frequently exclusively dictate the dominate instructional mode, instructional materials are major targets of teacher and community discontent (EPIE, 1977b)

In the choice of instructional materials school districts are often reliant upon subjective impressions of faculty, book publishers' pitches, and other school district or state authorities' recommendations. Among the curricula in the elementary school, social studies materials are particularly prone to criticism due to the lack of public agreement on core objectives and the controversial nature of social studies content. Social studies in the elementary school has drawn its content from the disciplines of history and geography more than other social sciences. Thus when a new curriculum was found which drew its content heavily from economics--though it addresses social systems problems that have long been standard fare in some form in the elementary curriculum (family roles, community interdependence, governmental structures and roles)--the curriculum committee felt a need to acquire firmer data for decision-making in recommending adoption or rejection of the social studies curriculum, Our Working World (Science Research Associates, 1973-74).

Methodology

The methodology involved designing a field test to evaluate the instructional materials against a set of social studies objectives which were accepted by the curriculum committee. This involved a number of steps, but the one of special concern was the development of a series of criterion referenced tests (CRT) which were to be used in the evaluation and for future testing in the social studies curriculum, providing it was adopted.

The curriculum did not come with prepackaged general objectives and the curriculum committee requested the publisher (SRA) to prepare these for each grade level. Twelve objectives were prepared for each grade 1 through 6. After examing these objectives and pronouncing them valid for their social studies program, the curriculum committee from the two districts requested the Office of Evaluation Research at the University of Illinois at Chicago

Circle to work with them in preparing items which would test student mastery of the concept in each of the twelve objectives.

A series of six tests were developed: for grades one and two there were 24 items per test, two per concept; for grades 3-6 there were 48 items per test, four per concept. These item limits were set in the interest of minimizing the testing time and building an instrument which could be administered at one sitting. A sample of three of the objectives and the items which were written to test them is given in Appendix A. It was soon discovered that objectives of this complexity pose unique difficulty in item writing for they do not lend themselves to neat learning hierarchies where mastery or competency can be defined as a behavior which is needed to move to the next higher level of instruction. In planning the tests three levels of refinement of items were used. First, a group of teachers at each level were asked to inspect the items and judge their suitability on two criteria: 1) readability and 2) accuracy in testing the concepts in the objectives. Second, three children at each grade level were given an individual administration of the test to check on readability. Third, a team of experts in social studies and elementary education screened the items for readability and accuracy in testing the concepts in the objectives.

The test for grade one was composed of pictoral items, the grade two test had mostly pictoral items, grade three had a few pictoral items and the other four grades used the conventional multiple choice and sequencing items.

Teachers were given written instructions on test administration. In the lower grades the test was read aloud by the teacher and was untimed at all grades. Teachers were cautioned not to interpret or explain any items for children.

Mastery levels for each of the grades was set by taking the mean of a series of judgments of experts, a system used by ETS in the Michigan State Assessment Tests. The mastery levels for each grade as set were:

Grade:	1	2	3	4	5	6
Mastery Level:	80%	75%	70%	70%	70%	70%

Eight classes were chosen at each grade level and were randomly assigned as control or experimental classes. The latter used the new curriculum Our Working World for social studies after the pre-test was administered. The control group used several more conventionally designed social studies programs that had been in use in the two districts. A pre-test was administered to all students in September before implementation of the social studies curricula and a post-test was administered in May after the experimental classes had been exposed to all the objectives. A total of 1,086 students took both the pre and post tests.

Two levels of analyses were run at each grade. The percent of students answering each item correctly and the percent of students mastering each concept; i.e., answering all the items (2 at grade level 1 and 2, and 3 of 4 at grade level 3-6) correctly was computed. For reasons to be discussed later, split-half reliabilities, an item discrimination index, and an item difficulty index were also computed.

Results

The results of the pre and post tests were first examined for the experimental and control group by item and by concept.

Table A below summarizes the mean percentage of experimental (E) and control (C) students answering the items correctly. The scores are calculated by taking the sum of the percentage of students answering each item correctly

(Σi) and dividing that by the number of items (N): $\frac{\Sigma i}{N}$ where Σi = % correct item 1 + % correct item 2 + . . . % correct item N.

Table A

Summary of Mean Percentage Item Mastery of Pre and Post Tests

Level	Experimental Pretest	Posttest	Gain	Control Pretest	Posttest	Gain
1.	69.29	80.04	10.75	73.20	79.62	6.42
2.	56.29	67.54	11.25	64.29	75.83	11.54
3.	65.81	77.10	11.29	63.85	71.79	7.94
4.	39.77	48.48	8.71	44.02	49.60	5.58
5.	47.79	53.60	5.81	45.67	49.60	3.93
6.	55.60	63.31	7.71	54.69	59.13	4.44

An examination of Table A shows that experimental students in Levels 1 and 3 reached specified mastery criteria on the posttest. For the control group, students in Levels 2 and 3 reached specified mastery cutoffs. More interesting is the amount of gain from the pretest to the posttest, also shown in Table A. The experimental group made higher gains for Levels 1, 3, 4, 5, and 6 and about the same at level 2 although the experimental group. started from a much lower level. These findings are graphed in Figure 1, which presents the amount of gain and Figure 2, which displays in graphic form pre and post mastery for both groups.

Table B reports the number of items for each level for which the experimental group students attained mastery levels on the posttest. For example, for Level 1, 17 items were mastered by 81 to 100% of the group, 4 items by 51 to 60%, and 3 items by 21 to 40%. (The reader is reminded that there are only 24 items for Levels 1 and 2, and 48 items for Levels 3--6).

Figure 1

School Districts 153 & 161 Social Studies Criterion Referenced Tests

Summary of Mean Percentage of Item Mastery Gain

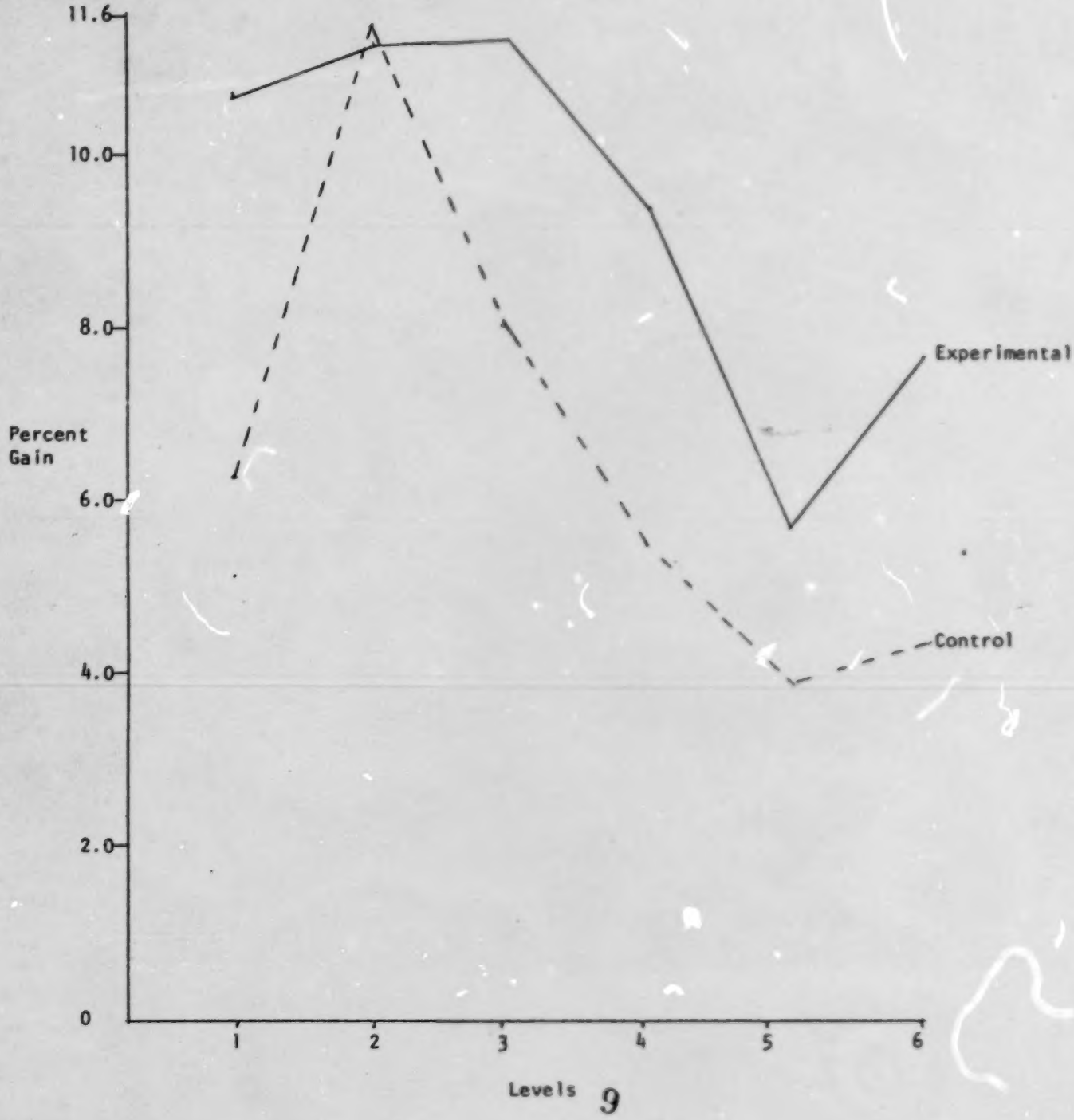

Figure 2

Social Studies Criterion Referenced Tests

Summary of Mean Percentage Item Mastery Pre and Post Test Scores

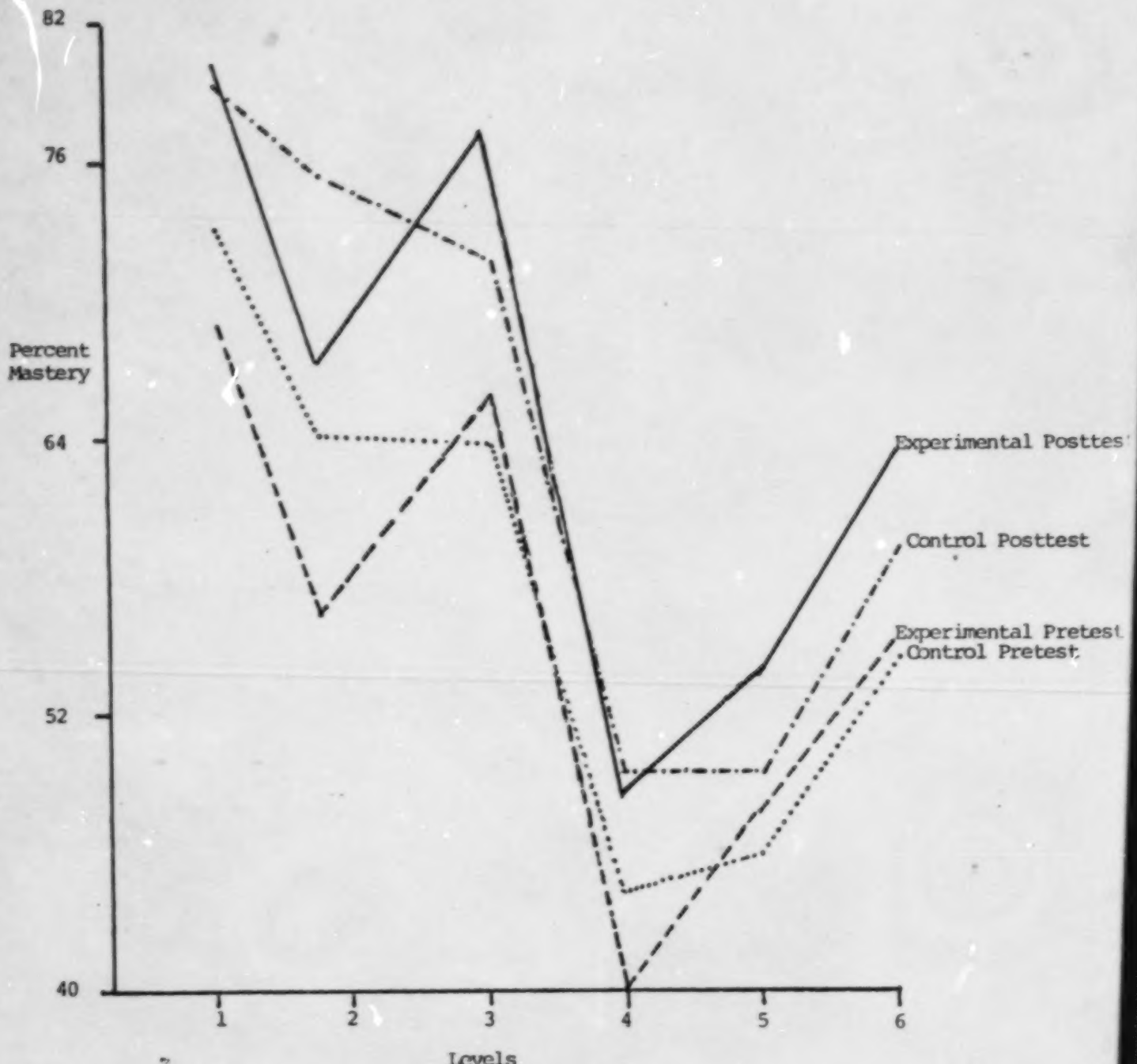

Table B

Breakdown of Percentage of Experimental Group Students Mastering Items

Level	Percentage of Students Mastering Item						
	0-20%	21-40%	41-50%	51-60%	61-70%	71-80%	81-100%
1	0	3	0	4	0	0	17
2	1	3	2	2	2	2	12
3	0	3	3	3	5	9	26
4	6	10	11	10	3	2	6
5	8	6	7	5	7	5	10
6	1	8	1	8	15	3	12

Mastery of concepts can be examined similarly. Table C presents information regarding the percentage of students mastering concepts.

Table C

Breakdown of Percentage of Experimental Group Students Mastering Concepts

Level	0-20%	21-40%	41-50%	51-60%	61-70%	71-80%	81-100%
1	0	3	0	3	0	0	6
2	2	3	1	2	1	3	0
3	0	1	1	1	1	3	5
4	3	5	2	0	2	0	0
5	3	4	1	2	1	1	0
6	2	1	1	2	2	3	1

Table C shows that for Levels 1, 2, 3 and 6 students approached mastery on at least half of the concepts.

For Level 5, mastery was incomplete for both experimental and control group students. For experimental students it is uncertain whether this is a product of problems with the test or if the objectives are not being learned

by students due to the absence of content or inadequate presentation. Therefore, no definite statement can be made by the evaluators although as we discuss later, we have reason to believe that the new curriculum content was difficult for teachers as well as children in the first teaching cycle. The fact that experimental students did make somewhat higher gains than did control students indicates that giving attention to the objectives and teaching could produce more positive results for experimental students next year.

The results for Level 4 indicate that the majority of concepts are not being mastered by experimental students. In the Level 4 analysis, it was found that the test items have high curriculum validity and good discriminating power. Therefore, it was recommended that the instructional practices in teaching Level 4 concepts be examined. Students may need greater exposure and may require more practice in using the concepts.

Finally the reliability of the post tests was computed. Listed below are the reliabilities calculated for both experimental and control groups of the post test for all 6 levels.

Table D

Summary Reliability of Total Tests

Level	1	2	3	4	5	6
Experimental	.71	.52	.76	.84	.77	.78
Control	.07	.63	.75	.80	.64	.77

The reliability of Levels 3, 4, 5 and 6 are in the satisfactory range. The results for Level 1 are mixed; while the reliability for the experimental group is certainly acceptable, especially considering the fact that the test is complete pictoral, the reliability for the control group, following general

test theory, is quite low. It is uncertain why the control reliability is so low. A great deal of guessing by control group students on the post test as the result of little or no exposure to the concepts might be one explanation. Since we are uncertain on the source of the low control reliability, we recommended that the reliability be checked again next year to see if problems with the test are indicated or if these results were idiosyncratic to this group. The reliability of Level 2 is not as high as would be preferred. A number of suggestions were given that might help increase reliability in the future, principally along the line of whether the content was being taught to children. The bulk of the report rendered to the school systems dealt with in-depth discussions of the results by level and their use in improving tests, curriculum and the instruction.

As originally requested by the school districts, the curriculum committee and classroom teachers received only the analysis discussed above, including the criterion referenced data or percentage of students mastering items and concepts by classroom. Following the delivery of the analysis report to the committee, OER - for its own information - examined the data further, using more sophisticated analyses including the computation of indexes of difficulty and indexes of discrimination for all items for both the experimental and control populations. After a brief period of study the committee returned to OER and requested more "statistical" information which they believed would be more "helpful" in making a decision on the curriculum and for revision of the tests. The data which had been prepared by OER for in-house only was given the districts and the final report was written incorporating both levels of analysis. Why did the curriculum committee and teachers change their minds when they had been such ardent advocates of criterion-referenced measurement at the beginning of the project? We believe their change of mind

when confronted with a need for a decision was an expression of some of the unresolved problems in CRT and indicates a narrower range of usefulness of CRT in curriculum decision-making than some of its enthusiasts have proclaimed.

Discussion

CRTs have been used in instructional sequences; they have not been used widely to evaluate materials for curriculum decision making. In their individual comparison approach they were considered by the curriculum committee to be a very attractive alternative when an evaluation of Our Working World was considered and NRT approaches were rejected. The remainder of the paper looks at the issues that emerged on CRT when used in the context of curriculum evaluation.

Criterion referenced tests (CRT) and curriculum adoption

CRTs have been used primarily to shape instruction for the individual learner. The commonly accepted definition of a CRT: "A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specific performance standards" (Glaser and Nitko, 1971) places an emphasis on individual performance against specific objectives. Thus objectives are written to contain specific criteria for the judgment of a learner's performance.

In the evaluation of a curriculum or program, however, the needed information is different; a distribution of variance in performance is desirable and needed. The individual learner question - can the student perform to criteria - is viewed as essentially a yes or no decision. As presented in the one widely accepted model of CRTs the approach does not provide the information that is needed by the decision maker to recommend adoption of a curriculum.

If a group of students performs to criteria then the curriculum may be too easy; i.e., students already have knowledge of the material and therefore do not need further instruction. If few perform to criteria what should be done - is the curriculum too difficult, or is the text at fault, are the objectives inappropriate, or have the performance objectives overshot the learning hierarchies necessary for their accomplishment? In any of these circumstances CRT gives limited and often not very useful information in aiding the curriculum decision maker. For example, CRT will not answer the question: If this curriculum is adopted, what provisioning costs are going to be incurred? What elements will have to be added by the teacher or school district to make it reasonably acceptable to large numbers of students of varying abilities?

A breakdown of the measures of achievement in NRTs does give more data on these questions. For example, an examination of the discrimination index on each item suggested that teachers' in-service was needed to emphasize the content to be taught and to supplement the content of the curriculum. Good learners as well as poor learners were mislead by distractors, giving evidence of lack of knowledge on content. We have good reason to suspect the low reliabilities on concepts in level 4 and 5 tests were functions of the teaching and not of the tests; i.e., students were not taught or given specific practice in the content and consequently gains were measures of increases in general education and not the curriculum. We came to this conclusion by checking the discrimination of items within the concepts and found that several did have good discrimination indexes as opposed to the total concept which had low reliability. In this case NRT data on reliability and discrimination gave more information related to the costs of provisioning if the curriculum was adopted than the CRT data on percentage of students obtaining mastery of each item or concept.

The use of CRTs may be inappropriate as a curriculum evaluation tool from another viewpoint. An implicit assumption in the items is that the content and behaviors sought by the curriculum are accurately identified. Hence the only problem to evaluate instruction is to see whether these goals are being mastered by the learner. Essentially the evaluation results are oriented to questions of time and methodology but not to whether the curriculum content is of sufficient scope and to whether it is related to significant social content.

Under this assumption major curriculum issues are sidestepped and the major weight of the evaluation is placed upon the vehicle of instruction. Social interests, social concerns and the interrelationship of concepts to values are more likely to be overlooked in the technology of evaluation employed.

Setting Mastery Levels

The literature reflects a range of views on setting levels of mastery and there is little theoretical agreement on what constitutes mastery. There seems to be a growing feeling that mastery levels stated in percentage figures are set arbitrarily and bear limited relationship to reality. At best they seem to be set to compensate for measurement error and student variability which cannot be squeezed out by a rigorous performance based model. In the attempt to escape from the dilemna of fixed performance standards in the State model, CRT based on a Continuum model (where mastery is a point on a continuum below which students will not be passed rather than a single standard of performance) is being advocated (Meskauhas, 1976). The curriculum committee did not find the setting of standards of performance useful inasmuch as it failed to provide direct information on how this curricula compared with other curriculum. The levels of performance were set with a floating

reference existing in the minds of "experts". The final decision on adoption rested on whether the experimental group gained more than the control group. Without the control group the customary way of presenting CRT data (number of students reaching mastery vs. non-mastery) would probably have resulted in a negative decision on adoption. Pre and post measures for the experimental group were made more meaningful when the control groups gains were introduced for comparison especially in those areas of formal social systems that Our Working World is especially designed to teach. Percentages of mastery arbitrarily set were discarded in the final decision-making on program adoption as they failed to bring any meaningful data forward on the comparative worth of curriculum and what is needed if students are to perform acceptably in the curriculum design of the new program. In short there is the large question of whether the setting of a percentage of mastery represents meaningful learning in a broad concept focused program. Our judgment was that it does not and, if used in program evaluation alone, may cause a program to be rejected on a basis that excludes other valuable learnings that would be made available if more adequate curriculum provisioning were arranged. The structuring of conclusions through different ways of reporting data on CRTs has been emphasized and concentrating on mastery of non-mastery can exaggerate or diminish the impact of the curriculum (Barta, et.al, 1976). The use of the fixed level of mastery in a curriculum evaluation which extends beyond skills is, we believe, a poor practice as it can lead to a premature judgment on the rejection of a program.

Macro-objectives and CRTs

Program evaluation in social studies is concerned with macro-objectives which embrace a cluster of behaviors and not a specific behavior as is found in skill performances. While there are specific skills in social studies, the

macro-objectives which are of special interest include a number of behaviors which have to be directed and synthesized in fashioning the totality of the learning. As an example, the objective from Level 3: "The student will explain how the city can be thought of as a system comprised of several interdependent parts" requires a synthesis of many learnings and not a specific isolated skill. As framed, the objectives produced a number of problems for the test developers which current models of CRT theory do not address. One response to this criterion might be to suggest that the macro-objective be broken down into a hierarchy of learning tasks which then are successively mastered (Gagne, 1967). Once the specific tasks are broken out, then the individual behaviors can be taught to and synthesized by the learner into the behavioral cluster that is involved in the citizenship understanding of a city as an interdependent system. Acting on this instructional design advice poses special problems for CRTs.

In breaking down the macro-objective, one quickly builds a lengthening list of behaviors each of which require a series of items if learner performance is to be directly assessed. An additional question then intrudes. If a student masters the series of specific items, will these add up into a behavioral pattern that is called into action when confronted with a problem that calls for, as in this case, an individual's analysis of a city as a system with interdependent parts? There is no evidence that behavioral patterns that result in solid citizens (presuming solid citizens are knowledgable about organized complexities) are developed through the learning of discrete behaviors that can be turned into test items easily amenable to the performance standards of a CRT. As a process which is excessively cumbersome the end result is a very lengthy series of tests to gauge learner performance and is not, in our judgment, scarcely an alternative. Moreover the process trivializes subject

matter as it atomizes it into discrete parts and raises fundamental questions on whether this is the mode for effective learning to take place. At the heart of this issue is the whole-part learning controversy. It is no small heresy to raise questions about the theory of learning being pursued, especially with rampant behaviorism now holding sway in curriculum and instruction, but for curriculum evaluation which purports to use CRT as its vehicle it is an issue which must be faced.

From this experience in social studies the domain of reference for test items cannot exclusively be specific skill oriented performance, especially when our interests are on macro-objectives which encompass broad behavioral patterns. This suggests that classical test theory with its NRTs may be a more significant anchor - one that is more comprehensible without excessive cost and measures citizenship behavior in a more readily interpretable context. Despite the seeming paradox, relative standards as expressed in NRTs of students' performance are more stable and socially more meaningful than CRTs based on the judgment of individual instructors in a limited item (1 to 4 per concept) approach. CRT theory has not addressed this issue, operating as it does on some implicit assumptions concerning learning. It will have serious limitations when employed in curriculum evaluation where macro-objectives are of primacy.

One technical advance which may be of use to curriculum evaluators in trying to bypass the problem of limited testing time, is suggested in the number of items theoretically that must be administered to obtain test reliability (Davis and Diamond, 1974, Millman, 1973). Their calculations suggest that if performance based objectives are to be used in curriculum evaluation, the number of items would have to be increased manyfold from the 2 to 4 per concept that were used in this study. If curriculum evaluators use the estimates of Davis and Diamond that a test of 20 items per objective should be used, then test

sampling by student with very large samples of students would be required for a curriculum evaluation, presuming that each performance specifically measured would require a test of twenty items. Economy of time and resources in curriculum evaluation usually imposes the restriction of a few items in CRTs to check mastery by the learner. Unfortunately the few items do not give sufficient range of a dimension of behavior which is the strongest single reason for employing the change to CRT. Our teachers were most uncomfortable with the CRT results as the enormity of the generalizing the mastery of the concepts rested on such slim item evidence. While test sampling has been used in one national curriculum evaluation (Walberg, 1970) it is probably not feasible at the district level for reasons of sample availability and cost.

In the social studies curriculum with broad macro-objectives, the interest of teachers proved to be in the range of the dimension of social behavior (in a broad sense) and obtaining the best description for understanding it within a domain. Can the dimension of social behavior best be understood as it is set forth in specific elements of subject matter which is tested by CRTs, (mastery-nonmastery), or is it best described by measures which are descriptive of its distribution within a population (NRT)? The curriculum committee, after examining both sets of data described previously, concluded that measures which are descriptive of population (or sample) distribution are most useful to curriculum adoption decision-making. These measures are the identifying mark of NRTs, which give a calibrated measure of the distribution of population characteristics. Woodson (1974) argues that variance in the test items is critical to providing useful information, and CRTs, in restricting this range in the items, may simply be limiting information. Which is more representative of the way we judge social behavior? Our evidence in this study suggests that a normative judgment is.

The absolute of the State model's all or none approach in the mastery-non-mastery CRT proved troublesome to teachers although as previously explained, this CRT model originally had appeal in avoiding comparisons of individuals and classrooms. The committee quickly discovered that, by examing CRT scores, the social behaviors in these tests are a matter of degree judged within a population and are not dichotomous nor ever accurately represented in a continuum model. Thus a composite test score that was placed within a known population was more informative than one which was established for the individual. The committee concluded that relative scores of NRTs are a more stable indicator of students' social learning than fixed absolute scores based on expert estimates, especially for purposes of curriculum adoption.

Evaluative Considerations

In the development of the tests the curriculum committee early on encountered a problem in the field testing. Field tests were conducted first with teachers and secondly with students to see if the tests' readability were suitable for the several levels. A copy of the tests and objectives were distributed to two teachers at each level for the purpose of obtaining their judgments on readability and curriculum validity. As a group they were very critical of the pictoral items and suggested many graphic as opposed to content changes. Fortunately the changes were not made prior to a field test with students; students evidenced no difficulty with the pictoral items on readability although they found the content difficult. This experience casts further doubts on the usefulness of mastery levels set a-priori by teachers as being creditable standards of judgment of pupil performance.

In this type of a curriculum evaluation the new material is at a great disadvantage as teachers are on the first cycle of teaching. A heavy burden is placed on the materials' instructional design as teachers learn the material

with the children. (In the above field test with teachers, teachers found many of the test items difficult and admitted they probably would not score well on the upper level tests). The scores obtained either as composites or on mastery of items may not be representative of what would be obtained in a second cycle of teaching the curriculum. Therefore in use of the findings generated for curriculum evaluation, they undoubtedly represent a conservative performance by students and teachers. However the results when analyzed by item discrimination, distractor counts and level of difficulty offered extensive information on curriculum and instruction provisioning needed if students were going to be successful. Of particular interest were the large gaps in students' knowledge that became apparent; e.g., a majority of students indicated that the chief executive of Illinois is the mayor. Extensive in-service suggestions for instructional time, technique and content were drawn for each level from the standard norm referenced statistics and the distractor counts which are given.

Conclusion

There is increasing interest in criterion-referenced testing as curriculum emphases shift to individual mastery of concepts from the traditional norm-referenced scores of standardized achievement testing. The value of criterion-referenced measurement and its relationship to classical test theory has been the subject of debate (Bernkopf and Bashau, 1976). How useful the information obtained through criterion-referenced measures in decision-making on selection of instructional materials is a question that has not been investigated as carefully as has the use of CRTs in guiding instruction. This study completed in forty-eight classrooms in two school districts found that teachers' predispositions toward criterion-referenced tests were weakened when they received the information of children's performance and were asked to make a decision on

adoption of instructional materials based on these results. They requested the traditional item analysis data and central tendency measures as well as the profiling of classes against the district's mean scores. Item discrimination scores and distractor counts were seen as particularly helpful by the curriculum committee. The school districts on the basis of the findings did adopt the new social studies curriculum. While criterion-referenced evaluation provided important information for aiding the school board in adopting a drastically different social studies program, it was not considered sufficient for making the adoption decision nor was it considered sufficient by the curriculum committee of administrators to direct inservice efforts. The field investigation found emerging at the decision-making level the issues of measurement that have emerged in the theoretical literature on CRTs. Because a decision was required there had to be a resolution of these issues. When a better data base for making adoptions and for directing an inservice program was sought, both approaches for analysis of the findings were employed. More importantly the CRT data are not an adequate substitute for NRT data in curriculum evaluation where a school district materials adoption decision is at stake.

References

Barta, M.B., Unhai, A.R., Gastright, J.F. Some problems in Interpreting Criterion-Referenced Test Results in A Program Evaluation. Studies in Educational Evaluation, 1976, 2, (3) 193-201.

Bernkopf, S. and Bashau, W. L. An Investigation of Criterion-Referenced Tests under Different Conditions of Sample Variability and Item Homogeneity. Paper presented at the meeting of the American Educational Research Association, San Francisco, April, 1976.

Carroll, J. B. A Model of School Learning. Teachers College Record, 1963, 64, 723-733.

Davis, F. B. and Diamond J. J. The Preparation of Criterion-Referenced Tests. In C. W. Harris et.al. (Eds.) Problems in Criterion-referenced Measurement. Los Angeles; UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

EPIE Report: Number 78. Selector's Guide for Elementary School Language Arts Programs. New York: EPIE Institute, 1977 (a).

EPIE Report: Number 76. Report on A National Study of the Nature and the Quality of Instructional Materials Most Used by Teacher and Learners. New York: EPIE Institute, 1977 (b).

Gagne, R. M. Curriculum Research and the Promotion of Learning. In R. W. Tyler et.al (Eds.) Perspectives of Curriculum Evaluation. Chicago. Rand McNally and Company, 1967, 19-38.

Glaser, G. R. and Nitko, A. J. "Measurement in Learning and Instruction." In R. L. Thorndike (Ed.) Educational Measurement (2nd Ed.) Washington, D.C. American Council on Education, 1971, 625-670.

Meskauskas, J. A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting. Review of Educational Research, 1976, 46 (1), 133-158

Millman, J. Passing Scores and Test Lengths for Domain-referenced Tests. Review of Educational Research, 1973, 43, 205-216.

Science Research Associates Our Working World, Lawrence Senesh. Series grades 1-6, 1973-74.

Walberg, Herbert "A Model for Research on Instruction" School Review, 1970, 78: 185-200.

Woodson, C. D. Classical Test Theory and Criterion-referenced Scales. N.D. (ERIC Document ED 083 298 filmed 2/13/74). 13p.

Appendix A

LEVEL ONE

Objective 12. The student should be able to cite examples of different attitudes and beliefs held by persons in their community.

23. If you could tell only one person about your visit to a baseball game, who do you feel would be most interested in listening ? Circle the picture of this one person.

YOUR DOCTOR

YOUR GYM TEACHER

A STORE CLERK

24. There are some buildings all people go to and there are other buildings that only some people go to. Circle the building to which only some people go.

SUPERMARKET

DEPARTMENT STORE

CHURCH

illustrated by C. Goldstein

Objective 12. The student will be able to make two lists, one citing direct and the other indirect costs of crime.

45-48. When a burglar steals from someone's house, there are direct and indirect costs of the burglary. A direct cost is the money spent because of the specific burglary. An indirect cost is the money spent to avoid future burglaries.

A burglar steals two TV sets, one radio, a record player, and many small items. Place an "X" on the line by four of the sentences that are examples of an Indirect Cost of this burglary.

____ THE FAMILY BUYS A NEW TV SET TO REPLACE THE STOLEN ONE.
____ THE FAMILY BUYS A NEW RADIO TO REPLACE THE STOLEN ONE.
____ THE FAMILY PUTS AN EXTRA LOCK ON THE DOOR.
____ MORE POLICEMEN ARE HIRED TO WATCH THE NEIGHBORHOOD.
____ A NEW JAIL IS BUILT IN TOWN.
____ A WATCH DOG IS BOUGHT BY THE FAMILY.
____ THE FAMILY SAVES MONEY BY NOT LEAVING FOR A VACATION.
____ THE FAMILY DISCOVERS A WEEK LATER THAT THE TOASTER IS MISSING.

LEVEL FIVE

Objective 12. The student will be able to cite several examples of episodes which challenged the existing social system and describe whether those challenges moved the system closer to or farther from the American ideals.

45-48. Choose the four events which have challenged the present social system and have moved the system closer to American ideals.

_____ government can record private conversations of citizens

_____ giving everyone a lawyer when he/she is arrested for a crime

_____ stopping the printing of news that criticizes politicians

_____ parents refusing to send children to school

_____ giving everyone an equal chance to qualify for a job

_____ giving everyone the right to vote

_____ having only one major political party

_____ government protecting the right of every citizen to purchase or rent a house in any community